

Protein databases

Henrik Nielsen

CENTER FOR
RBIOL
CAL SEQU
ENCE ANA
LYSIS CBS

Swiss Institute of Bioinformatics (SIB) and European Bioinformatics Institute (EBI)

National Biomedical Research Foundation, Georgetown University, USA

In 2002 merged into:

A collaboration between SIB, EBI and Georgetown University.



UniProt

UniProt Knowledgebase (UniProtKB)

UniProt Reference Clusters (UniRef)

UniProt Archive (UniParc)

UniProt Metagenomic and Environmental Sequence
Database (UniMES)

UniProt Knowledgebase Release 2013_02 (06-Feb-13)
consists of:

UniProtKB/Swiss-Prot: Annotated manually (*curated*)

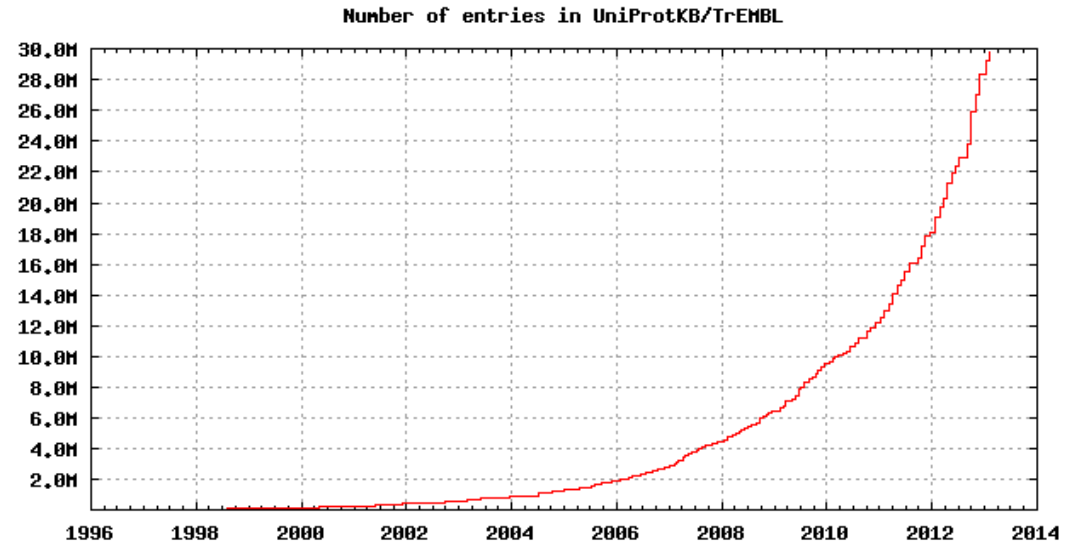
539,165 entries

UniProtKB/TrEMBL: Computer annotated

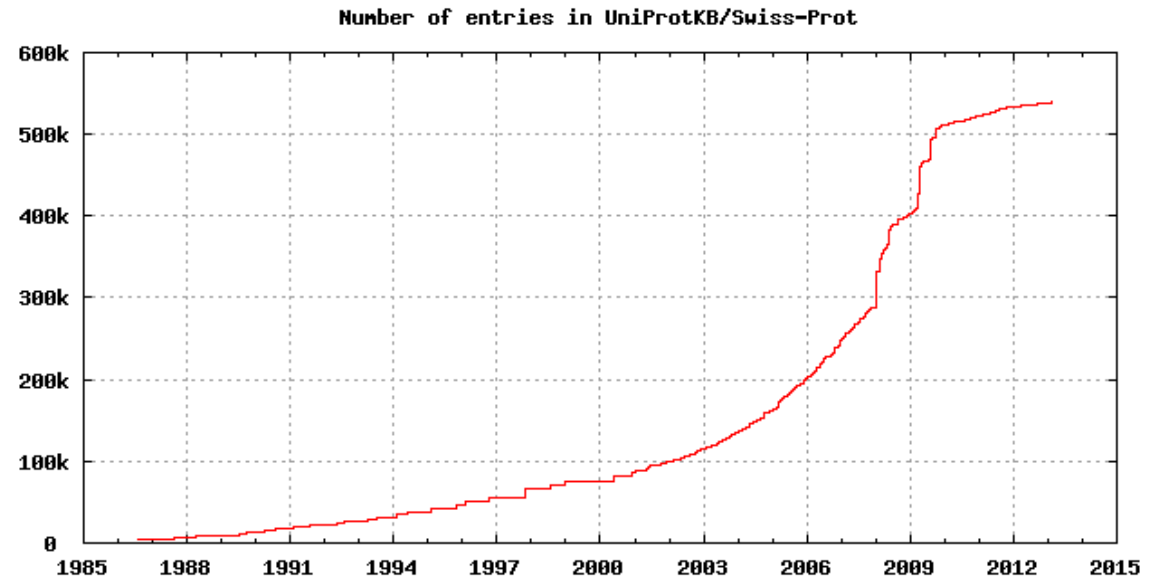
29,769,971 entries

Growth of UniProt

TrEMBL



Swiss-Prot



Content of UniProt Knowledgebase

- Amino acid sequences
 - Functional and structural annotations
 - Function / activity
 - Secondary structure
 - Subcellular location
 - Mutations, phenotypes
 - Post-translational modifications
 - Origin
 - organism: Species, subspecies; classification
 - tissue
 - References
 - Cross references
-

Amino acid sequences

From where do you get amino acid sequences?

- Translation of nucleotide sequences (GenBank/EMBL/DDBJ)
 - Direct amino acid sequencing: *Edman degradation*
 - Mass spectrometry
 - 3D-structures
-

UniProt entry, formatted view

UniProt > UniProtKB

Downloads · Contact · Documentation/Help

Search Blast * Align * Retrieve ID Mapping *

Search in Protein Knowledgebase (UniProtKB) Query Search Advanced Search > Clear

P01009 A1AT HUMAN Reviewed, UniProtKB/Swiss-Prot
Last modified January 25, 2012. Version 180. History...

Contribute
Send feedback
Read comments (0) or add your own

Clusters with 100%, 90%, 50% identity Documents (6) Third-party data text xml rdf/xml gff fasta

Names · Attributes · General annotation · Ontologies · Interactions · Alt products · Sequence annotation · Sequences · References · Web links
Cross-refs · Entry info · Documents · Customize order

Names and origin

| | |
|----------------------|---|
| Protein names | <p>Recommended name: Alpha-1-antitrypsin</p> <p>Alternative name(s): Alpha-1 protease inhibitor Alpha-1-antiproteinase Serpine A1</p> <p>Cleaved into the following chain: 1. Short peptide from AAT Short name=SPAAT</p> |
| Gene names | <p>Name: SERPINA1</p> <p>Synonyms: AAT, PI</p> <p>ORF Names: PRO0684, PRO2209</p> |
| Organism | Homo sapiens (Human) |
| Taxonomic identifier | 9606 [NCBI] |
| Taxonomic lineage | Eukaryota > Metazoa > Chordata > Craniata > Vertebrata > Euteleostomi > Mammalia > Eutheria > Euarchontoglires > Primates > Haplorrhini > Catarrhini > Hominidae > Homo |

Protein attributes

UniProt entry, text view (flat file)

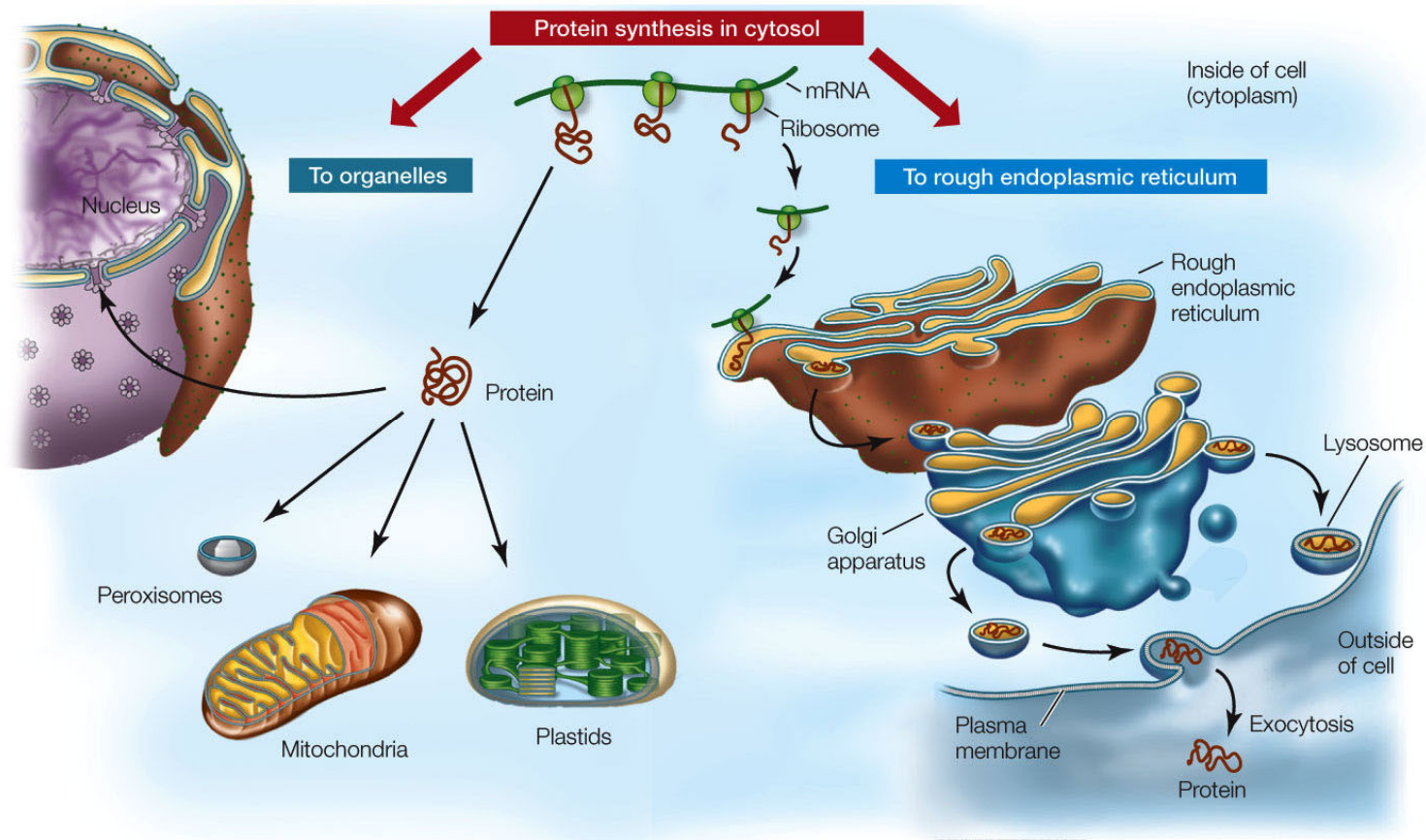
```
ID      A1AT_HUMAN                      Reviewed:          418 AA.
AC      P01009; A6PX14; B2RDQ8; QOPVP5; Q13672; Q53XB8; Q5UOM1; Q7M4R2;
AC      Q86U18; Q86U19; Q96BF9; Q96ES1; Q9P1P0; Q9UCE6; Q9UCM3;
DT      21-JUL-1986, integrated into UniProtKB/Swiss-Prot.
DT      01-OCT-1996, sequence version 3.
DT      25-JAN-2012, entry version 180.
DE      RecName: Full=Alpha-1-antitrypsin;
DE      AltName: Full=Alpha-1 protease inhibitor;
DE      AltName: Full=Alpha-1-antiprotease;
DE      AltName: Full=Serpine A1;
DE      Contains:
DE          RecName: Full=Short peptide from AAT;
DE              Short=SPAAT;
DE      Flags: Precursor;
GN      Name=SERPINA1; Synonyms=AAT, PI; ORFNames=PRO0684, PRO2209;
OS      Homo sapiens (Human).
OC      Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC      Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
OC      Catarrhini; Hominidae; Homo.
OX      NCBI_TaxID=9606;
RN      [1]
RP      NUCLEOTIDE SEQUENCE [MRNA] (ISOFORM 1).
RX      MEDLINE=84107980; PubMed=6319097;
RA      Bollen A., Herzog A., Cravador A., Herion P., Chuchana P.,
RA      van der Straten A., Loriau R., Jacobs P., van Elsen A.;
RT      "Cloning and expression in Escherichia coli of full-length
RT      complementary DNA coding for human alpha 1-antitrypsin.";
RL      DNA 2:255-264(1983).
...
```

General annotation (Comments)

General annotation (Comments)

| | |
|---------------------------------|--|
| Function | <p>Inhibitor of serine proteases. Its primary target is elastase, but it also has a moderate affinity for plasmin and thrombin. Irreversibly inhibits trypsin, chymotrypsin and plasminogen activator. The aberrant form inhibits insulin-induced NO synthesis in platelets, decreases coagulation time and has proteolytic activity against insulin and plasmin. Ref.17 Ref.18 Ref.24</p> <p>Short peptide from AAT (SPAAT) is a reversible chymotrypsin inhibitor. It also inhibits elastase, but not trypsin. Its major physiological function is the protection of the lower respiratory tract against proteolytic destruction by human leukocyte elastase (HLE). Ref.17 Ref.18 Ref.24</p> |
| Subcellular location | <p>Secreted Ref.24.</p> <p>Short peptide from AAT: Secreted > extracellular space > extracellular matrix Ref.24.</p> |
| Tissue specificity | Plasma. |
| Domain | The reactive center loop (RCL) extends out from the body of the protein and directs binding to the target protease. The protease cleaves the serpin at the reactive site within the RCL, establishing a covalent linkage between the carboxyl group of the serpin reactive site and the serine hydroxyl of the protease. The resulting inactive serpin-protease complex is highly stable. |
| Post-translational modification | <p>Several isomers are observed, resulting from the combination of different N-linked glycan structures and mature N-terminus. N-linked glycan at Asn-107 is alternatively di-antennary, tri-antennary or tetra-antennary, whereas glycan at Asn-70 is di-antennary with trace amounts of tri-antennary, and glycan at Asn-271 is exclusively di-antennary. The structure of the antennae is Neu5Ac(alpha1-6)Gal(beta1-4)GlcNAc attached to the core structure Man(alpha1-6)[Man(alpha1-3)]Man(beta1-4)GlcNAc(beta1-4)GlcNAc. Some antennae are fucosylated, which forms a Lewis-X determinant. Proteolytic processing may yield the truncated form that ranges from Asp-30 to Lys-418.</p> |
| Polymorphism | The sequence shown is that of the M1V allele which is the most common form of PI (44 to 49%). Other frequent alleles are: M1A 20 to 23%; M2 10 to 11%; M3 14 to 19%. |
| Involvement in disease | <p>Defects in SERPINA1 are the cause of alpha-1-antitrypsin deficiency (A1ATD) [MIM:613490]. A disorder whose most common manifestation is emphysema, which becomes evident by the third to fourth decade. A less common manifestation of the deficiency is liver disease, which occurs in children and adults, and may result in cirrhosis and liver failure. Environmental factors, particularly cigarette smoking, greatly increase the risk of emphysema at an earlier age. Ref.58 Ref.60 Ref.62</p> |
| Miscellaneous | The aberrant form is found in the plasma of chronic smokers, and persists after smoking is ceased. It can still be found ten years after smoking has ceased. |
| Sequence similarities | Belongs to the serpin family . |
| Sequence caution | <p>The sequence CAD62334.1 differs from that shown. Reason: Erroneous initiation. Translation N-terminally shortened.</p> <p>The sequence CAD62585.1 differs from that shown. Reason: Erroneous initiation. Translation N-terminally shortened.</p> |

Protein sorting in eukaryotes



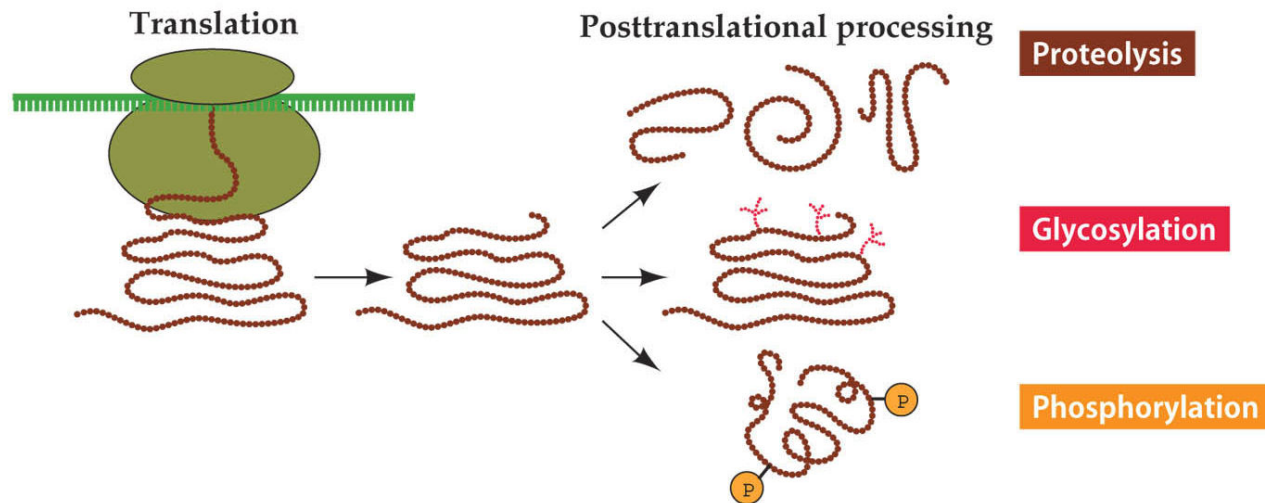
Different proteins belong to different compartments of the cell – and some belong *outside* the cell

General annotation (Comments)

General annotation (Comments)

| | |
|---------------------------------|--|
| Function | <p>Inhibitor of serine proteases. Its primary target is elastase, but it also has a moderate affinity for plasmin and thrombin. Irreversibly inhibits trypsin, chymotrypsin and plasminogen activator. The aberrant form inhibits insulin-induced NO synthesis in platelets, decreases coagulation time and has proteolytic activity against insulin and plasmin. Ref.17 Ref.18 Ref.24</p> <p>Short peptide from AAT (SPAAT) is a reversible chymotrypsin inhibitor. It also inhibits elastase, but not trypsin. Its major physiological function is the protection of the lower respiratory tract against proteolytic destruction by human leukocyte elastase (HLE). Ref.17 Ref.18 Ref.24</p> |
| Subcellular location | <p>Secreted Ref.24.</p> <p>Short peptide from AAT: Secreted > extracellular space > extracellular matrix Ref.24.</p> |
| Tissue specificity | Plasma. |
| Domain | The reactive center loop (RCL) extends out from the body of the protein and directs binding to the target protease. The protease cleaves the serpin at the reactive site within the RCL, establishing a covalent linkage between the carboxyl group of the serpin reactive site and the serine hydroxyl of the protease. The resulting inactive serpin-protease complex is highly stable. |
| Post-translational modification | <p>Several isomers are observed, resulting from the combination of different N-linked glycan structures and mature N-terminus. N-linked glycan at Asn-107 is alternatively di-antennary, tri-antennary or tetra-antennary, whereas glycan at Asn-70 is di-antennary with trace amounts of tri-antennary, and glycan at Asn-271 is exclusively di-antennary. The structure of the antennae is Neu5Ac(alpha1-6)Gal(beta1-4)GlcNAc attached to the core structure Man(alpha1-6)[Man(alpha1-3)]Man(beta1-4)GlcNAc(beta1-4)GlcNAc. Some antennae are fucosylated, which forms a Lewis-X determinant. Proteolytic processing may yield the truncated form that ranges from Asp-30 to Lys-418.</p> |
| Polymorphism | The sequence shown is that of the M1V allele which is the most common form of PI (44 to 49%). Other frequent alleles are: M1A 20 to 23%; M2 10 to 11%; M3 14 to 19%. |
| Involvement in disease | <p>Defects in SERPINA1 are the cause of alpha-1-antitrypsin deficiency (A1ATD) [MIM:613490]. A disorder whose most common manifestation is emphysema, which becomes evident by the third to fourth decade. A less common manifestation of the deficiency is liver disease, which occurs in children and adults, and may result in cirrhosis and liver failure. Environmental factors, particularly cigarette smoking, greatly increase the risk of emphysema at an earlier age. Ref.58 Ref.60 Ref.62</p> |
| Miscellaneous | The aberrant form is found in the plasma of chronic smokers, and persists after smoking is ceased. It can still be found ten years after smoking has ceased. |
| Sequence similarities | Belongs to the serpin family . |
| Sequence caution | <p>The sequence CAD62334.1 differs from that shown. Reason: Erroneous initiation. Translation N-terminally shortened.</p> <p>The sequence CAD62585.1 differs from that shown. Reason: Erroneous initiation. Translation N-terminally shortened.</p> |

Post-translational modifications



Many proteins are *modified* after their synthesis in order to become active

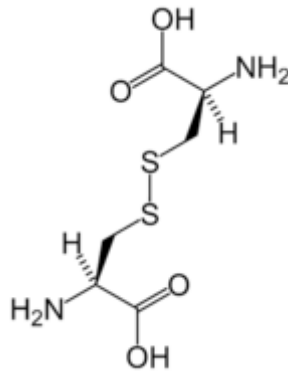
Proteolysis: Cleavage of *signal peptides*, *propeptides* or *initiator methionine*

Glycosylation: Particularly common in proteins on the *surface* of cells. Also plays a role in sorting of proteins to *lysosomes*

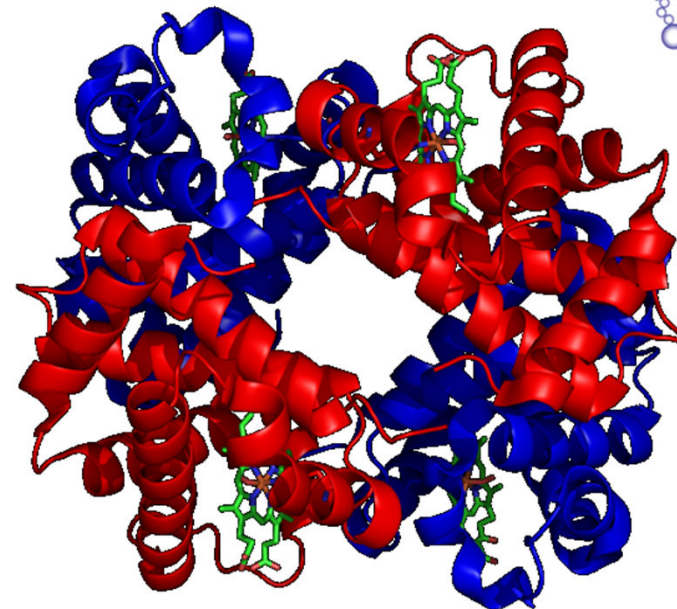
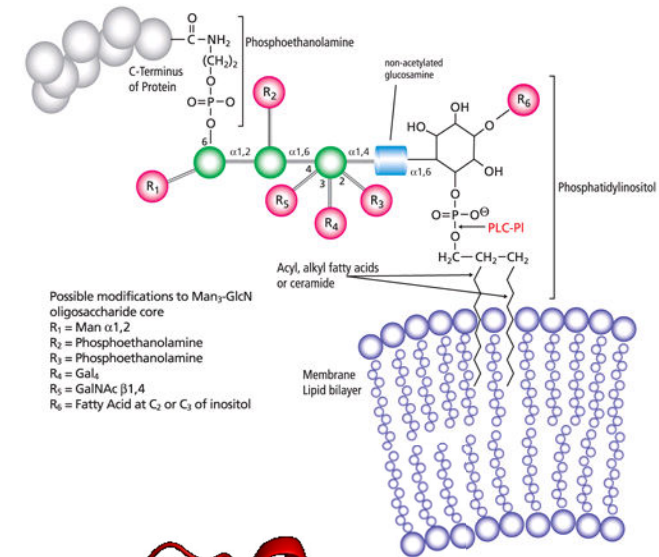
Phosphorylation: Often *reversible*. Regulates the *activity* of many enzymes

More post-translational modifications

- Lipid anchors
 - (e.g. GPI anchors)
- Disulfide bonds



- Prosthetic groups
 - (e.g. metal ions)



General annotation (Ontologies)

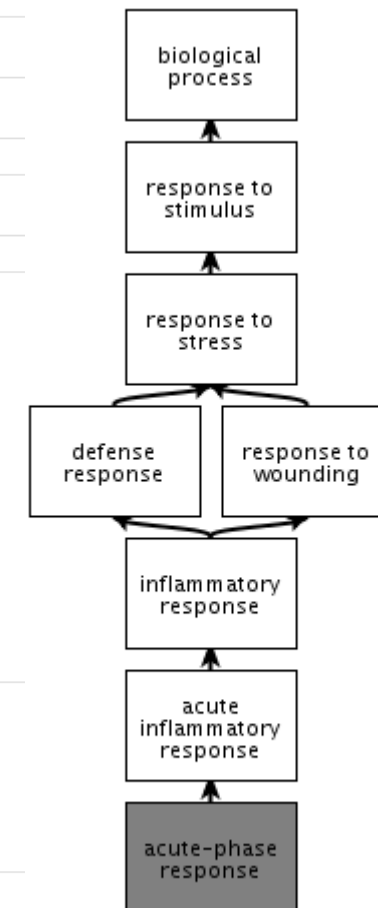
Ontologies

Keywords

| | |
|---------------------------|--|
| Biological process | Acute phase Blood coagulation |
| Cellular component | Extracellular matrix Secreted |
| Coding sequence diversity | Alternative splicing Polymorphism |
| Domain | Signal |
| Molecular function | Protease inhibitor Serine protease inhibitor |
| PTM | Glycoprotein |
| Technical term | 3D-structure Complete proteome Direct protein sequencing Reference proteome |

Gene Ontology (GO)

| | |
|--------------------|--|
| Biological process | <p>acute-phase response Inferred from electronic annotation. Source: UniProtKB-KW</p> <p>platelet activation Traceable author statement. Source: Reactome</p> <p>platelet degranulation Traceable author statement. Source: Reactome</p> <p>regulation of proteolysis Inferred from Biological aspect of Ancestor. Source: RefGenome</p> |
| Cellular component | <p>extracellular space Inferred from mutant phenotype. Source: UniProtKB</p> <p>platelet alpha granule lumen Traceable author statement. Source: Reactome</p> <p>proteinaceous extracellular matrix Inferred from electronic annotation. Source: UniProtKB-SubCell</p> |
| Molecular function | <p>protease binding Inferred from physical interaction. Source: UniProtKB</p> <p>serine-type endopeptidase inhibitor activity Non-traceable author statement. Source: UniProtKB</p> |



QuickGO - <http://www.ebi.ac.uk/QuickGO>

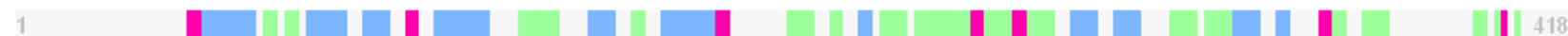
Sequence annotation (Feature Table)

Sequence annotation (Features)

| | Feature key | Position(s) | Length | Description | Graphical view | Feature identifier |
|---------------------------------|----------------------|-------------|--------|---|----------------|--------------------|
| Molecule processing | | | | | | |
| <input type="checkbox"/> | Signal peptide | 1 – 24 | 24 | Ref.15 Ref.16 Ref.17 Ref.18 | | |
| <input type="checkbox"/> | Chain | 25 – 418 | 394 | Alpha-1-antitrypsin Ref.2 | | PRO_0000032377 |
| <input type="checkbox"/> | Peptide | 375 – 418 | 44 | Short peptide from AAT | | PRO_0000364030 |
| Regions | | | | | | |
| <input type="checkbox"/> | Region | 368 – 392 | 25 | RCL | | |
| Sites | | | | | | |
| <input type="checkbox"/> | Site | 382 – 383 | 2 | Reactive bond | | |
| Amino acid modifications | | | | | | |
| <input type="checkbox"/> | Modified residue | 256 | 1 | S-cysteinyln cysteine | | |
| <input type="checkbox"/> | Glycosylation | 70 | 1 | N-linked (GlcNAc...) (complex) Ref.19 Ref.26 Ref.27 Ref.28 Ref.29 Ref.30 Ref.31 Ref.32 | | |
| <input type="checkbox"/> | Glycosylation | 107 | 1 | N-linked (GlcNAc...) (complex) Ref.19 Ref.29 Ref.31 Ref.32 | | |
| <input type="checkbox"/> | Glycosylation | 271 | 1 | N-linked (GlcNAc...) (complex) Ref.19 Ref.26 Ref.27 Ref.28 Ref.29 Ref.31 Ref.32 | | |
| Natural variations | | | | | | |
| <input type="checkbox"/> | Alternative sequence | 307 – 418 | 112 | Missing in isoform 3. | | VSP_028890 |
| <input type="checkbox"/> | Alternative sequence | 356 – 418 | 63 | AVHKA...NPTQK → VRSP in isoform 2. | | VSP_028889 |
| <input type="checkbox"/> | Natural variant | 4 | 1 | S → L in Z-Wrexham. Ref.62 | | VAR_006978 |
| <input type="checkbox"/> | Natural variant | 26 | 1 | D → A in V-Munich. Ref.69 | | VAR_006979 |
| <input type="checkbox"/> | Natural variant | 37 | 1 | T → A. Corresponds to variant rs11558262 [dbSNP Ensembl]. | | VAR_051938 |
| <input type="checkbox"/> | Natural variant | 58 | 1 | A → T in M5-Karlsruhe. | | VAR_006980 |
| <input type="checkbox"/> | Natural variant | 63 | 1 | R → C in I. Ref.63 Corresponds to variant rs28931570 [dbSNP Ensembl]. | | VAR_006981 |

Secondary structure (Feature Table)

Secondary structure



Helix Strand Turn

[Details...](#)

| | | | | | |
|--------------------------|-------------|-----------|----|--|--|
| <input type="checkbox"/> | Turn | 48 – 50 | 3 | | |
| <input type="checkbox"/> | Helix | 51 – 68 | 18 | | |
| <input type="checkbox"/> | Beta strand | 70 – 72 | 3 | | |
| <input type="checkbox"/> | Beta strand | 74 – 76 | 3 | | |
| <input type="checkbox"/> | Helix | 78 – 89 | 12 | | |
| <input type="checkbox"/> | Helix | 94 – 103 | 10 | | |
| <input type="checkbox"/> | Turn | 108 – 110 | 3 | | |
| <input type="checkbox"/> | Helix | 113 – 127 | 15 | | |
| <input type="checkbox"/> | Beta strand | 135 – 145 | 11 | | |
| <input type="checkbox"/> | Helix | 152 – 160 | 9 | | |
| <input type="checkbox"/> | Beta strand | 165 – 169 | 5 | | |
| <input type="checkbox"/> | Helix | 174 – 188 | 15 | | |
| <input type="checkbox"/> | Turn | 189 – 191 | 3 | | |
| <input type="checkbox"/> | Beta strand | 206 – 215 | 10 | | |
| <input type="checkbox"/> | Beta strand | 218 – 220 | 3 | | |
| <input type="checkbox"/> | Helix | 224 – 226 | 3 | | |
| <input type="checkbox"/> | Beta strand | 228 – 237 | 10 | | |
| <input type="checkbox"/> | Beta strand | 239 – 256 | 18 | | |
| <input type="checkbox"/> | Turn | 257 – 260 | 4 | | |
| <input type="checkbox"/> | Beta strand | 261 – 268 | 8 | | |
| <input type="checkbox"/> | Turn | 269 – 271 | 3 | | |
| <input type="checkbox"/> | Beta strand | 272 – 279 | 8 | | |

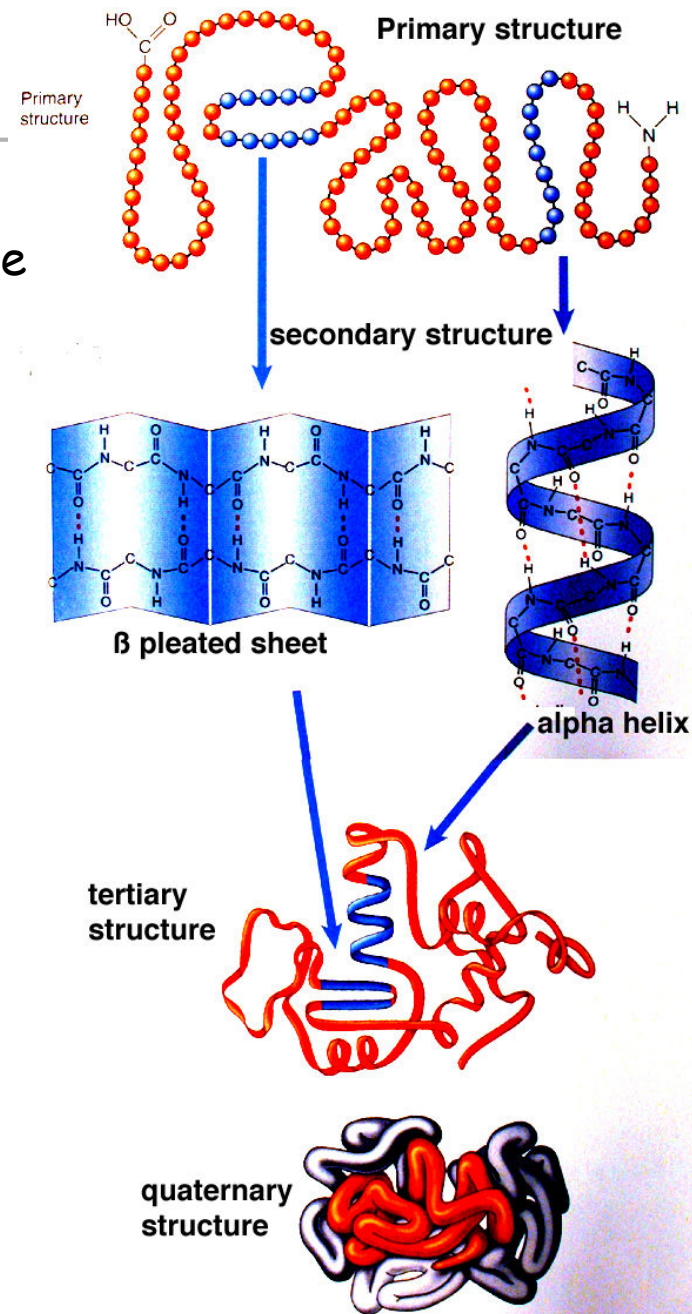
Protein structure

Primary structure: Amino acid sequence

Secondary structure:
"Backbone" hydrogen bonding
Alpha helix / Beta sheet / Turn

Tertiary structure: Fold, 3D coordinates

Quaternary structure: subunits



Evidence (Comments, Feature Table)

Q43495 (108_SOLLC) ★ Reviewed, UniProtKB/Swiss-Prot

Last modified March 2, 2010. Version 49. [History...](#)

General annotation (Comments)

| | |
|-----------------------|---|
| Subcellular location | Secreted Potential |
| Tissue specificity | Stamen- and tapetum-specific. |
| Sequence similarities | Belongs to the A9/FIL1 family . |

Sequence annotation (Features)

| Feature key | Position(s) | Length | Description | Graphical view | Feature identifier |
|---|-------------|--------|----------------------------|----------------|--------------------|
| Molecule processing | | | | | |
| <input type="checkbox"/> Signal peptide | 1 – 30 | 30 | Potential | | |
| <input type="checkbox"/> Chain | 31 – 102 | 72 | Protein 108 | | PRO_0000000238 |
| Amino acid modifications | | | | | |
| <input type="checkbox"/> Disulfide bond | 41 ↔ 77 | | By similarity | | |
| <input type="checkbox"/> Disulfide bond | 51 ↔ 66 | | By similarity | | |
| <input type="checkbox"/> Disulfide bond | 67 ↔ 92 | | By similarity | | |
| <input type="checkbox"/> Disulfide bond | 79 ↔ 99 | | By similarity | | |

Evidence/Confidence types

- 3 types of *non-experimental qualifiers* in
Sequence annotation and General comment:
- *Potential*: Predicted using sequence analysis
 - *Probable*: Uncertain experimental evidence
 - *By similarity*: Predicted using sequence similarity
-

UniProt entry, sequence(s)

Sequences

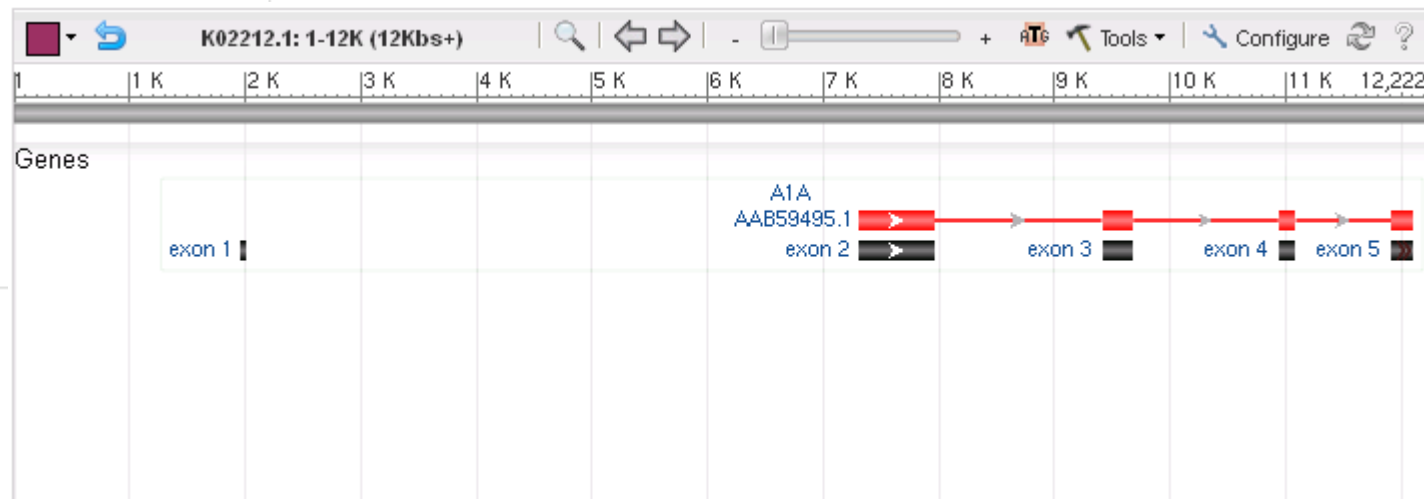
| Sequence | | Length | Mass (Da) | Tools |
|--|-------|--------|-----------|---|
| <input type="checkbox"/> Isoform 1 [UniParc]. | FASTA | 418 | 46,737 | Blast <input type="button" value="go"/> |
| Last modified October 1, 1996. Version 3. Checksum: 7016555F273B7F16 | | | | |
| <div><div>102030405060</div><div>MPSSVSWGIL LLAGLCCLVP VSLAEDPQGD AAQKTDTS HH DQDHPTFNKI TPNLAEF AFS</div><div>708090100110120</div><div>LYRQLAHQSN STNIFFSPVS IATAFAMLSL GTKADTHDEI LEGLNFNLTE IPEAQIHEGF</div><div>130140150160170180</div><div>QELLRTL NQP DSQLQLTTGM GLFLSEGLKL VDKFLEDVKK LYHSEAF TVN FGDTEEAKKQ</div><div>190200210220230240</div><div>INDYVEKGTQ GKIVDLVKEL DRD TVFALVN YIFFKGKWER PFEVKDTEEE DFHVDQVTTV</div><div>250260270280290300</div><div>KVPMMKRLGM FNIQHCKKLS SWVLLMKYLG NATAIFFLPD EGKLQHLENE LTHDIITKFL</div><div>310320330340350360</div><div>ENEDRRSASL HLPKLSITGT YDLKSVLGQL GITKVFSNGA DLSGVTEEAP LKLSKAVHKA</div><div>370380390400410</div><div>VLTI DEKGTE AAGAMFLEAI PMSIPPEVKF NKPFVFLMIE QNTKSPLFMG KVVNPTQK</div></div> | | | | |
| « Hide | | | | |
| <input type="checkbox"/> Isoform 2 [UniParc]. | FASTA | 359 | 40,263 | Blast <input type="button" value="go"/> |
| Checksum: D16A255538FB2945 Show » | | | | |
| <input type="checkbox"/> Isoform 3 [UniParc]. | FASTA | 306 | 34,755 | Blast <input type="button" value="go"/> |
| Checksum: 15C708E6C25CE0C4 Show » | | | | |

Cross-references, nucleotide sequences

Sequence databases

- ☒ EMBL
- ☐ GenBank
- ☐ DDBJ

[K01396](#) mRNA. Translation: [AAB59375.1](#).
[K02212](#) Genomic DNA. Translation: [AAB59495.1](#).
[X01683](#) mRNA. Translation: [CAA25838.1](#).
[M11465](#) mRNA. Translation: [AAA51546.1](#).
[J02619](#) Genomic DNA. Translation: [AAA51547.1](#).
[DQ682455](#) mRNA. Translation: [ABG73380.1](#).
[AM048838](#) Genomic DNA. Translation: [CAJ15161.1](#).
[AF113676](#) mRNA. Translation: [AAF29581.1](#).
[AF130068](#) mRNA. Translation: [AAG35496.1](#).
[BX161449](#) mRNA. Translation: [CAD61914.1](#).
[BX247968](#) mRNA. Translation: [CAD62306.1](#).
[BX248002](#) mRNA. Translation: [CAD62334.1](#). Different initiation.
[BX248257](#) mRNA. Translation: [CAD62585.1](#). Different initiation.
[AK315637](#) mRNA. Translation: [BAG38005.1](#).
[BT019455](#) mRNA. Translation: [AAV38262.1](#).

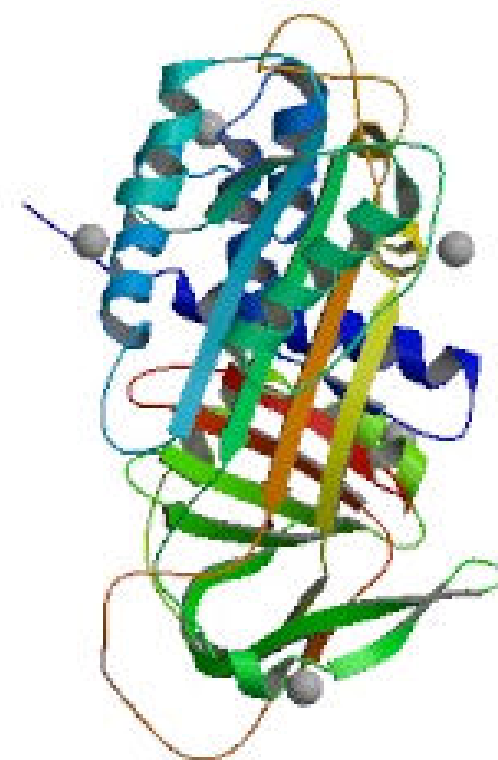


Cross-references, 3D structure

3D structure databases

- PDBe
- RCSB PDB
- PDBj

| Entry | Method | Resolution (Å) | Chain | Positions | PDBsum |
|-------|--------|----------------|-------|-----------|--------|
| 1ATU | X-ray | 2.70 | A | 45-418 | [>] |
| 1D5S | X-ray | 3.00 | A | 44-377 | [>] |
| | | | B | 378-418 | [>] |
| 1EZK | X-ray | 2.60 | A | 48-382 | [>] |
| | | | B | 383-418 | [>] |
| 1HP7 | X-ray | 2.10 | A | 25-418 | [>] |
| 1IZ2 | X-ray | 2.20 | A | 25-418 | [>] |
| 1KCT | X-ray | 3.46 | A | 25-418 | [>] |
| 1O08 | X-ray | 2.65 | A | 26-418 | [>] |
| 1OPH | X-ray | 2.30 | A | 26-418 | [>] |
| 1PSI | X-ray | 2.92 | A | 26-418 | [>] |
| 1QLP | X-ray | 2.00 | A | 26-418 | [>] |
| 1QMB | X-ray | 2.60 | A | 49-376 | [>] |
| | | | B | 377-418 | [>] |
| 2D26 | X-ray | 3.30 | A | 26-382 | [>] |
| | | | B | 383-418 | [>] |
| 2QUG | X-ray | 2.00 | A | 25-418 | [>] |
| 3CWL | X-ray | 2.44 | A | 25-418 | [>] |
| 3CWM | X-ray | 2.51 | A | 25-418 | [>] |
| 3DRM | X-ray | 2.20 | A | 26-418 | [>] |
| 3DRU | X-ray | 3.20 | A/B/C | 26-418 | [>] |
| 3NDD | X-ray | 1.50 | A | 46-372 | [>] |
| | | | B | 383-418 | [>] |
| 3NDF | X-ray | 2.70 | A | 46-381 | [>] |
| | | | B | 383-418 | [>] |
| 3T1P | X-ray | 3.90 | A | 48-418 | [>] |
| 7API | X-ray | 3.00 | A | 36-382 | [>] |
| | | | B | 383-418 | [>] |
| 8API | X-ray | 3.10 | A | 36-382 | [>] |
| | | | B | 383-418 | [>] |
| 9API | X-ray | 3.00 | A | 36-382 | [>] |
| | | | B | 383-418 | [>] |



Cross-references

Other databases linked from UniProt

(there are ~100 in total):

- Nucleotide sequences
 - 3D structure
 - Protein-protein interactions
 - Enzymatic activities and pathways
 - Gene expression (microarrays and 2D-PAGE)
 - Ontologies
 - Families and domains
 - Organism specific databases
-

Translation and Reading Frames

The genetic code

| | | Second letter | | | | |
|--------------|---|--|------------------------------------|--|---|------------------|
| | | U | C | A | G | |
| First letter | U | UUU Phenyl-alanine UUC UUA Leucine UUG | UCU UCC Serine UCA UCG | UAU Tyrosine UAC UAA Stop codon UAG Stop codon | UGU Cysteine UGC UGA Stop codon UGG Tryptophan | U C A G |
| | C | CUU CUC Leucine CUA CUG | CCU CCC Proline CCA CCG | CAU Histidine CAC CAA Glutamine CAG | CGU CGC Arginine CGA CGG | U C A G |
| | A | AUU AUC Isoleucine AUA AUG Methionine; start codon | ACU ACC Threonine ACA ACG | AAU Asparagine AAC AAA Lysine AAG | AGU Serine AGC AGA Arginine AGG | U C A G |
| | G | GUU GUC Valine GUA GUG | GCU GCC Alanine GCA GCG | GAU Aspartic acid GAC GAA Glutamic acid GAG | GGU GGC Glycine GGA GGG | U C A G |

- Degenerate (*redundant*) but not ambiguous
- *Almost* universal (deviations found in mitochondria)

Reading Frames 1

A piece of an mRNA-strand:

5' aug cccaagcugaaauagcguagagggguuuucaucauuugaggacgauguaaa 3'

can be divided into triplets (*codons*) in three ways:

| | | | | | | | | | | | | | | | | | | |
|---|--|-----|-----|--------------------------------------|--------------------------------------|-----|--------------------------------------|-----|-----|-----|-----|-----|--------------------------------------|-----|-----|-----|-----|--------------------------------------|
| 1 | aug | ccc | aag | cug | aa | agc | gua | gag | ggg | uuu | uca | uca | uuu | gag | gac | gau | gua | uaa |
| | M | P | K | L | N | S | V | E | G | F | S | S | F | E | D | D | V | * |
| 2 | ugc | cca | agc | uga | aua | gcg | uag | agg | ggg | uuu | cau | cau | uug | agg | acg | aug | uau | |
| | C | P | S | * | I | A | * | R | G | F | H | H | L | R | T | M | Y | |
| 3 | gcc | caa | gcu | gaa | uag | cgu | aga | ggg | guu | uuc | auc | auu | uga | gga | cga | ugu | aua | |
| | A | Q | A | E | * | R | R | G | V | F | I | I | * | G | R | C | I | |

Each possible set of triplets is called a *reading frame*.

Reading Frames 2

Since there are two strands in DNA, there are *six* possible reading frames in a piece of DNA (three in each direction):

| | | | | | | | | | | | | | | | | | | | |
|----|---|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----|----|
| 3 | A | Q | A | E | * | R | R | G | V | F | I | I | * | G | R | C | I | | |
| 2 | C | P | S | * | I | A | * | R | G | F | H | H | L | R | T | <u>M</u> | <u>Y</u> | | |
| 1 | <u>M</u> | <u>P</u> | <u>K</u> | <u>L</u> | <u>N</u> | <u>S</u> | <u>V</u> | <u>E</u> | <u>G</u> | <u>F</u> | <u>S</u> | <u>S</u> | <u>F</u> | <u>E</u> | <u>D</u> | <u>D</u> | <u>V</u> | * | |
| 5' | ATGCCCAAGCTGAATAGCGTAGAGGGGTTTTTCATCATTTGAGGACGATGTATAA | | | | | | | | | | | | | | | | | 3' | |
| 3' | TACGGGTTCGACTTATCGCATCTCCCCAAAAGTAGTAAACTCCTGCTACATATT | | | | | | | | | | | | | | | | | 5' | |
| | H | G | L | Q | I | A | Y | L | P | K | * | * | K | L | V | I | Y | L | -1 |
| | | G | L | S | F | L | T | S | P | N | E | D | N | S | S | S | T | Y | -2 |
| | <u>A</u> | <u>W</u> | <u>A</u> | <u>S</u> | <u>Y</u> | <u>R</u> | <u>L</u> | <u>P</u> | <u>T</u> | <u>K</u> | <u>M</u> | <u>M</u> | Q | P | R | H | I | -3 | |

A reading frame from a start codon to the first stop codon is called an *open* reading frame (underlined above).

Virtual Ribosome — required reading!

Nucleic Acids Research, 2006, Vol. 34, Web Server issue W385–W388
doi:10.1093/nar/gkl252

Virtual Ribosome—a comprehensive DNA translation tool with support for integration of sequence feature annotation

Rasmus Wernersson*

Center for Biological Sequence Analysis, BioCentrum-DTU, Technical University of Denmark,
Building 208, DK-2800 Lyngby, Denmark

Received February 14, 2006; Revised March 1, 2006; Accepted March 20, 2006

ABSTRACT

Virtual Ribosome is a DNA translation tool with two areas of focus. (i) Providing a strong translation tool in its own right, with an integrated ORF finder, full support for the IUPAC degenerate DNA alphabet and all translation tables defined by the NCBI taxonomy group, including the use of alternative start codons. (ii) Integration of sequences feature annotation—in particular, native support for working with files containing intron/exon structure annotation. The software is available for both download and online use at <http://www.cbs.dtu.dk/services/VirtualRibosome/>.

This makes it easy to build datasets that can be used for analyzing how the underlying exon structure is reflected in the protein [e.g. how exon modules maps onto the 3D structure of the protein, see the FeatureMap3D server (4) elsewhere in this issue].

SOFTWARE FEATURES

Support for the degenerate nucleotide alphabet

The software has full support for the IUPAC alphabet (Table 1) for degenerate nucleotides. For example, the codon TCN correctly translates to S (serine) and not X (unknown) as often seen in other translators.

Support for a wide range of translation tables

Full support for all translation tables defined by the NCBI taxonomy group (5) (see the list below). The command-line version of the software also has support for

INTRODUCTION

A large number of software packages for translating DNA sequences already exist, as services on the World Wide